JUN 0 8 1988

ED 1.310/2:285893

COMPLETED
18

ED285893

Center for Student Testing, Evaluation
and Standards

DELIVERABLE - NOVEMBER 1986

TESTING STUDY GROUP: The Impact of Advances in Artificial Intelligence on Test Development

Evaluation Approaches to Intelligent Computer-Assisted Instruction

Study Director: Eva L. Baker

Grant Number: OERI-G-86-0003

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

TM 870 470

ORIGINAL

The project presented or reported herein was partially performed pursuant to a grant from the Office of Educational Research and improvement/Department of Education (OERI/ED). However, the opinions expressed herein do not necessarily reflect the position or policy of the OERI/ED and no official endorsement by the OERI/ED should be inferred.

DOCUMENT RESUME

ED 285 893 TM 870 470

AUTHOR Baker, Eva L.
TITLE Evaluation Approaches to Intelligent Computer-Assisted Instruction. Testing Study Group: The Impact of Advances in Artificial Intelligence on Test Development.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Nov 86
GRANT OERI-G-86-0003
NOTE 18p.
PUB TYPE Viewpoints (120) -- Reports - General (140)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Artificial Intelligence; *Computer Assisted Instruction; Courseware; Criterion Referenced Tests; *Evaluation Methods; Evaluation Problems; Evaluators; *Formative Evaluation; Measurement Objectives; Program Development; Program Evaluation; *Summative Evaluation
IDENTIFIERS *Domain Referenced Tests; *Intelligent CAI Systems

ABSTRACT

Some special problems associated with evaluating intelligent computer-assisted instruction (ICAI) programs are addressed. This paper intends to describe alternative approaches to the assessment and improvement of such applications and to provide examples of efforts undertaken and shortfalls. Issues discussed stem chiefly from the technical demands of the artificial intelligence field, which have tended to limit most evaluation efforts to first-party evaluation by project staff. ICAI evaluation should make use of a range of formative (e.g., componential analysis) and summative (e.g., cost analysis) evaluation methods with multiple criterion measures. Standardized tests have not proved sensitive enough in this area; domain-referenced tests are especially well suited to ICAI, because their success depends on experts' care in constructing detailed specifications of the knowledge domain. Individual differences in students' intelligence, cognitive styles, and state anxiety should also be considered. As evaluators sharpen their goals, they will be able to select the most relevant data to collect, and present useful instructional options. Special recommendations for ICAI evaluation include: (1) developing an expectation of evaluation; (2) rewarding evaluation participation; (3) increasing credibility of the evaluating team by encouraging expert participation; (4) adapting evaluation to specific features of ICAI development; (5) performing componential analysis of software under development; and (6) maintaining both a responsible and responsive approach. (LPG)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Evaluation Approaches to Intelligent Computer-Assisted Instruction

Eva L. Baker
University of California, Los Angeles
Harold F. O'Neil, Jr.
University of Southern California

Evaluation is an activity purported to provide an improved basis for decision-making. Among its key elements are the identification of goals, the assessment of process, the collection of information, analysis, and the interpretation of findings. The most common model for evaluation is summative (Scriven, 1967) which focuses on overall choices among systems or programs based upon performance levels, time, and cost. In this mode, evaluation is essentially comparative and contrasts the innovation against other options. These comparisons may be against explicit choices or may be implicit in terms of current practice or ways resources might be spent in the future (opportunity costs).

Evaluation efforts that are instituted at the outset or in the process of an innovation's development typically have different purpose. Formative evaluation (Baker, 1973) seeks to provide information that focuses on the improvement the innovation and is designed to assist the developer. Formative evaluation also addresses, from a metaevaluation perspective, the effectiveness of the development procedures used, in order to predict whether the application of similar approaches will likely have effective, and efficient, results. In that function, formative evaluation seeks to

improve the technology at large rather than the specific instances addressed one at a time.

A critical issue in any sort of evaluation is the meaning ascribed to the findings. Meaning derives from the use of measures that are valid for the intervention, from the adequacy of the inferencing processes used to interpret results, and from the utility of the findings for the intended users. These facets of meaning require that the designer/developer as well as funding sources articulate their goals, processes, and potential decision needs so that the evaluation team can provide results that have meaning for interested parties.

Tensions in Evaluation

A persistent fact of evaluation is that those evaluated rarely see the value of the process. It is something done to them, a necessary evil, a new chance for failure, often seen as largely irrelevant to their major purpose. This view generally holds whether it is a person who is evaluated (for selection or credentialling purposes), such as students and teachers at universities or in the public schools, a program evaluated (either as small as a segment of curriculum or as large as a federal initiative), or a technological innovation. Those who get evaluated are almost always recalcitrant players.

As persistent a fact, however, is that those in authority have come to believe that evaluation is a useful process. Their belief is fostered in part by actual research studies (Weiss, 1975; Alkin, 1984) that evaluation findings, when used, improve the state of affairs. But more a likely reason that evaluation has been fastened upon as a useful endeavor resides

in the belief that it provides a mechanism for management, or for the appearance of management, by those in charge of resources. Objectivity, accountability, and efficiency are themes underlying this commitment to evaluation.

The tension is obvious between those who are reluctant participants and those who push the evaluation process frompositions of authority. Evaluation experts have to mediate among these two sets of views, a challenging if not always fun role.

This chapter is addressed to the evaluation of new technology, specifically intelligent computer-assisted instruction (ICAI) applications. We intend to describe alternative approaches to the assessment and improvement of such applications and to provide examples of efforts undertaken and their shortfalls, as well as to sketch alternatives.

The Evaluability of ICAI Applications

Evaluating an emerging technology presents serious technical as well as practical problems, and the ICAI field incorporates most known or imaginable difficulties. First, much has been claimed by proponents of AI. The claims have led many sponsors to support projects that are nominally addressed to the development of an innovation (such as a tutor), but in fact the intentions of the designers is not to create a working, effective tutor, but to explore the limits of the field. Rather, in this case, the tutor is a context, a constraint under which the designer really seeks to conduct research, i.e., produce new knowledge about AI processes. If the tutor happens, it is more than a side effect, but less than the

major goal. Thus, the lines between research and application are murky and undercut neat categories of R & D processes, such as those identified by Glennan (1965) and Bright (1970) and used as program elements in DoD work (6.1, 6.2, etc.). This reality presents problems for evaluation. Compared to other innovations, the ICAI what to be evaluated is less concrete and identifiable, and more like the probabilistic view of where a photon is at any point in time.

Secondly, the public persona of AI, (see national magazines, films, television, trade books) is high profile. In startling contrast, the accessibility to AI processes is limited. To the uninitiated, it is embedded in the recesses of special language. Coupled with the fact that AI work is conducted at a relatively few centers by a relatively small number of people, understanding an AI implementation well enough to create sensible options for its assessment is difficult proposition. These states are compounded by the strongly capalistic environment in which AI reseasrch is conducted. The proprietary nature of much work, either that conducted by large private corporations (Xerox or BBN, for example) or by small entrepeneurial enterprises also works to obscure the conceptual and procedural features of the work. AI applications are unlike, therefore, innovations in health, criminal justice, education, industrial training, employment, transportation because of the lack of mid-level communication about what is actually is. Perhaps AI experts can assist in evaluation, but, understandably, they are more interested in creating something new of their own. All of this is asserted with full knowledge that at least some of these problems characterize any rapidly developing new technology.

The utility of evaluation processes also needs to be judged in terms of what techniques and options are useful, where there is differential confidence in our ability to measure and inference, and which procedures have been used credibility in the last ten years. In addition, we must consider what requirements ICAI evaluation creates and explore new methodology to meet these needs.

In the civilian sector, over the last ten years increasing use of an evaluation approach. The approach, formative evaluation, is designed so that its principal outputs are identification of success and failure of segments, components, and details of programs rather than a simple overall estimate of project success. The approach requires that data be developed to permit the isolation of elements for improvement and ideally, the generation of remedial options to assure that subsequent revisions have a higher probability of success. Formative evaluation is a method that developed to assist in the development of instructional (training) programs. While the evaluation team maintains "third-party" objectivity, they typically interact with and understand program goals, processes, and constraints at a deeper level than evaluation teams focused exclusively on bottom line assessments of success or failure. Their intent is to assist their client (either funding agency or project staff) to use systematic data collection to promote the improvement of the effort. Basic literature in formative evaluation was developed by Scriven (1967); Baker and Alkin (1972); Baker (1973); Baker and Saloutos (1974). Formative evaluation now represents the major focus of evaluation efforts in the public education sector (Center for the Study of Evaluation, 1985) and permits the

integration of a variety of quantitative and qualitative data collection and analysis techniques to meet the goal of program improvement.

The Distance Between the Evaluator and the Evaluated

One way to think about evaluation techniques is in terms of the distance among those who are conducting the evaluation work, those responsible for the actual day-to-day design and development of the project, and those who are responsible for providing resources to the project. These distances are ofter represented as the "party" of the evaluation.

First party evaluation is evaluation conducted by the project staff itself. Common examples would be pilot test data conducted for input into the design of a final product. It has the benefit of intimate connection and understanding of the project. Its problem is lack of distance and detachment. In AI applications, this evaluation work is informal, and, perhaps relatively infrequently addressed to the issue of overall effectiveness of the intervention.

Second party evaluation involves the assessment of progress or outcomes by the supervising or funding agency. IPRs and site visits are examples of second party evaluation. Arbitrary timing, limited attention spans, and objectivity is a problem here.

Third party evaluation is evaluation conducted by an independent group. GAO performs some third party evaluation. Independent contractors reporting to state legislatures, school boards or school districts also conduct such evaluation. The benefit of such an approach is the disinterested nature of the investigation, contributing to the credibility of the findings.

However, the validity of external evaluation presents come difficulty, and requires, however, that the third party get up to speed on technical issues so that the evaluation methodologies applied are appropriate. The learning required by the evaluation staff represents an additional "overhead" to the project staff and may be perceived as a distraction fro their primary effort. This sort of evaluation costs more than the other two.

All types of above evaluation can be done using formative or summative techniques. Formative techniques are to develop information useful to the funding agency and to the project staff for improving the effectiveness of the effort. Summative evaluation is the standard experimental design control group approach used for go-no-go decisions. Third party evaluation is often summative, e.g., GAO.

Contrary to popular practice, there is no inherent reason for totally separating formative and summative evaluation efforts. The approaches differ in purpose and client. They also differ in the types of data appropriate (cost for summative; componential analysis for formative). However, in the area of performance they should share some common criterion measures. Failing that, the construct validity of multiple criterion measures should be established.

The remainder of this chapter will deal with aproaches to formative and summative evaluation for ICAI applications, and particularly the problems of external validity and generalizability of results.

Evaluation Technology

It is not likely that evaluation as it is currently practiced can be transferred directly to an application field such as ICAI. The issue

should be the marginal benefits for applying or adapting such technology to the new area of development.

One approach to exploring the merging of existing technologies (ICAI applications with evaluation technology) is to shift points of view in order to determine where reasonable matches exist. Looking first from the evaluation perspective, let us explore where evaluation has some strengths and could make substantial contribution to ICAI development.

Research and development in measurement is one of the major productive areas in psychology. Sophisticated models for estimating performance have been developed and come in and out of vogue. Many of these were created to assist in the selection process, to sort those individuals who were better or worse with regard to a particular competency or academic domain. These approaches, while venerable, have little to contribute to the assessment of programs, either those completed or under continuing development. Most standardized achievement tests are based on this model, and their use to evaluate innovation is not recommended for a variety of technical reasons. These reasons can be summed up in a simple phrase: standardized tests are not sensitive enough to particular curriculum foci, thus they are unlikely to detect effects present (the false negative problem) and will underestimate effects that exist.

Measurement of Student Achievement Outcomes

However, there are newer approaches to the measurement of human performance which do have implications for the assessment of ICAI interventions designed to improve learner performance. Specifically, the use of domain-referenced achievement testing seems to provide a good match

with ICAI approaches. In domain-referenced testing (Hively, Patterson, Page, 1968; Baker & Herman, 1983; Baker & O'Neil, 1986) one attempts to estimate student performance in a well specified content domain. The approach is essentially top-down, with parameters for content selection and criteria for judging adequacy of student output specified (albeit successively revised) in advance. Test items are conceived as samples from a universe constrained by the specified parameters. For example, in the area of reading comprehension, parameters would need to be explicated regarding the genre and content to be read, the characteristics of the semantics and syntax, including variety, ambiguity, complexity of sentence patterns, and the presupposed knowledge that the learner would bring into the instructional/testing setting. In addition, the characteristics of the items would be identified, in terms of gross format, i.e., short answer, essay, multiple choice, and in terms of subtler features such as the rules for the construction of wrong answer alternatives, or for the assessment of free responses. Theoretically, such rules permit the generation of a universe of test items which can be multiple re-sampled to provide progress and end-of- instruction testing.

The use of such approaches have the added benefit of utility to small numbers of students. They do not depend as do the selection approach described above upon normal (and large) distributions of respondents to derive score meaning. On the other hand, such tests are more demanding to develop, and they depend upon close interaction with the innovation designer, to assure that the specifications are adequate. They contrast with the common approach of "tacking on" existing measures (like

commercially available standardized tests), an easy enough process but one unlikely to provide information useful for the fair assessment of improvement or a product. Domain-referenced tests derive their power from the goodness of their specifications. Their weakness is their idiosyncracy; however, the matching of testing procedures to designer's intentions is also their strength.

Because of the attention that ICAI applications devote to representing properly the knowledge domain and determining student understanding in process, the application of improved assessment techniques, particularly those based on domain-referenced testing seems like a good fit.

Measurement of Individual Difference

A second area in measurement that could contribute to the efficient design and assessment of ICAI applications is the measurement of individual differences. Psychology has long invested resources in determining how best to assess constructs along which individuals show persisting differences. For these areas to be useful, such constructs should interact (statistically) with instructional options and desired outcomes of the system under study. Common constructs such as ability and intelligence undoubtedly have relevance for the analysis and implementation of alternative student models and tutoring strategies. Other constructs related to cognitive style preferences, e.g., the need for structure, the need for reflection, the attribution of success and failure, could illuminate design options and results analyses for ICAI applications. Similarly, constructs related to affective states, i.e., state anxiety (O'Neil, 1978) could also provide explanations of findings otherwise obscure.

Process Measurement and Analysis

In formative evaluation, much is made of the role of process evaluation, that is, tracking what occurs when to assure that inferences about system effectiveness are well placed. Central to this function, however, is deciding, to the extent possible, what data should be collected and which inferences should be drawn from the findings. Technology-based innovations often make two seemingly conflicting classes of errors. One error is collecting everything possible that can be tracked. Student response times, system operation, errors, student requests, etc. can be accumulated ad nauseum. The facts seem to be that rarely do developers attend to this glut of information. They have no strategies for determining how such data should be sorted, arranged in priority, nor ways to draw systematic conclusions from findings. By the time the database is assembled, developers are often on to new ideas and prospects; old data, particularly painfully analyzed and interpreted old data, remain only old, and often unused.

The other error in technology process measurement is when relevant information which could be painlessly accumulated and tabulated on-line is ignored.

The challenge for the evaluator is to help decide what data are likely to be most relevant. Relevance will presuppose a clear overall goal, such as teaching a target group a set of skills. In fact, in the entire gamut of measurement options available, the most significant contribution evaluators' may make is clarifying the goals that the designer possesses but has not articulated. Because of the mixture of research and

development goals inherent in much ICAI work in education, this is a nontrivial problem. The designer may feel he/she has all the goals that can be tolerated.

Generation of Instructional Options

Formative evaluators can assist designers to explore different ways in which they can successfully meet their goals. Of particular interest, for example, is the extent to which evaluation can highlight alternatives for the instructional strategies used in the application. In all instructional development, not the least in ICAI-based approaches, the designer fastens early upon a particular strategy. Research findings have suggested that teachers and developers are most reluctant to change the approach they have taken. They will play at the edges rather than rethink their method (Baker, Herman, & Yeh, 1981). Furthermore, they could easily adapt their basic approach by adding particular instructional options to their basic plan, assuming that they make their choice informed by prior research. In a recent study (Baker, Bradley, Aschbacher, & Feifer, 1985) WEST was experimentally modified to strengthen its teaching capability. That study was part of an effort designed to influence through formative evaluation the instructional design of AI tutors. The study was largely unsuccessful because of timing issues, but the concept remains to be adequately tested.

How Can Evaluation Assist ICAI Applications: An Agenda

The history of evaluation of AI implementations is a thin book. It consists of studies comparing expert systems and expert people. General student outcomes (like a standardized math achievement test) have been used to assess very specific and relatively short term interventions, usually to

no avail. At the heart of both of these approaches is comparisons, an idea not particularly conducive to the open flow of information in a competitive environment. For evaluation to work to the mutual benefit of application designers and their resource providers it must meet certain criteria.

1. The expectation of evaluation should be developed. The description of effectiveness of applications needs to become part of the socialized ethic, as in science, the ethic of repeatability, verifiability and public reporting is commonplace.

2. Rewards for evaluation participation are necessary. These must be over and above the intrinsic value of the evaluation information for the designer since evaluation is not a common expectation, special benefits must be developed to create cooperation.

3. The credibility of evaluation team must be seriously addressed. Interesting AI experts in this field (we have had a few notable successes) needs to depend less on frantic persuasion and more on a developed sense of professional responsibility (like reviewing for a journal). If the approach taken is formative, then the designer can receive "help" from friendly reviewers. The goal of evaluation of this sort is to aid in revision rather than render a judgment.

4. Approaches to evaluation must take account of specific features of ICAI development. Rather than waiting for the completed development, the evaluation team can assist in some decision-making related to instruction or utilization. While this sounds easy, it depends upon the view that "outsiders" know psychology or performance measurement in ways that may be useful to ICAI experts. We need to overcome the "not invented here" syndrome.

5. Evaluation needs to be componential and focus on the utility of the piece of software under development. Records of rapid prototyping and redesign need to be integrated into the formative evaluation. It is as useful to record the blind alleys as the successes (at least as far as the rest of the technology field is concerned).

6. Evaluation needs to be responsible and responsive. Objectivity must be preserved, but at the same time, those evaluated must not feel victimized. A reasonably positive example occurred in the formative evaluation of PROUST (Baker, Fiefer, & Aschbacher, 1985). Among the most interesting phases of that activity was the dialog following the submission of the draft of the report to Soloway. Through an extended process, the evaluation report was strengthened, fuller understanding of the intentions and accomplishments of the project staff developed, and points of legitimate disagreement identified. In all cases, the AI expert was able to present (directly quoted) his point of view. The overall outcome was that the fairness of the report was not questioned.

As cost issues become more salient during the next few years, evaluation pressures will rise. They are highly correlated. Anticipating how the field can deal with these pressures now, rather than waiting for the surprise to occur seems smart. To the extent it wants, the ICAI field can control the next direction of evaluation.

References

Baker, E.L. The technology of instructional development (Chapter 8). In R.M.W. Travers (Ed.), The second handbook of research on teaching. Chicago: Rand McNally, 1973. pp. 245-285.

Baker, E.L., & Alkin, M.C. Formative Evaluation of Instructional Development. AV Communication Review, 21(4), 1973.

Baker, E.L., Bradley, C., Aschbacher, P., & Feifer, R. An evaluation of WEST: An ICAI program. Los Angeles: UCLA Center for the Study of Evaluation, 1985.

Baker, E.L., Feifer, R., & Aschbacher, P.. An evaluation of PROUST: An ICAI program. Los Angeles: UCLA Center for the Study of Evaluation, 1985.

Baker, E.L., & Herman, J.L. Task Structure Design: Beyond Linkage. Journal of Educational Measurement, 1983, 20, 149-164.

Baker, E., Herman, J., & Yeh, J. Fun and games: Their relationship to basic skills in elementary school children. American Educational Research Journal, 1981, 18(1), 83-92.

Baker, E.L., & O'Neil, H.F., Jr. Assessing instructional outcomes. In Gagne, R. (Ed.), Instructional technology. Hillsdale, NJ: Lea/Wiley, in press.

Baker, E.L., & Soloutos, W.A. Formative evaluation of instruction. Los Angeles: UCLA Center for the Study of Evaluation, 1974.

Baker, E.L., & O'Neil, H.F., Jr. "Evaluation Approaches to Intelligent Computer-Assisted Instruction." Invited presentation at the U.S. Office of Personnel Management's Workshop on Potential Applications of Intelligent Computer Assisted Instruction in Military Training. Los Angeles, California, March, 1986.

Center for the Study of Evaluation. Center for Student Testing, Evaluation and Standards: Assessing and improving educational quality. Proposal submitted to OERI, August 1985. Los Angeles: UCLA Center for the Study of Evaluation.

Glennan, T.K., Jr. Issues in the choice of development policies. In T. Marshall, T.K. Glennan, & R. Summers (Eds.), Strategies for research and development. New York: Springer Velaz, 1965.

Hively, W., Patterson, H.L., & Page, S. A universe defined system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

O'Neil, H.F., Jr. (Ed.). Learning strategies. New York: Academic Press, 1979.

Scriven, M. Aspects of curriculum development. In R. Tyler (Ed.), Perspectives of curriculum evaluation. Chicago: Rand McNally, 1967.

BEST COPY AVAILABLE